Transportation Research Part A 100 (2017) 135–151

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Coauthorship network in transportation research

Lijun Sun [a,*], Iyad Rahwan [a,b]

[a] The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Institute for Data, Systems & Society, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form 8 April 2017
Accepted 12 April 2017
Available online 26 April 2017

Keywords:
Coauthorship network
Transportation research
Scientific collaboration
Social network analysis

ABSTRACT

The field of transportation research has been accelerating in the last decade. In this paper, we examine the structure of scientific collaboration in transportation research by building a coauthorship network using publication metadata from 22 transportation journals. In this network, a vertex represents a researcher and an edge represents the collaboration (coauthorship) between a pair of researchers. To build an accurate network, we propose and apply an efficient author name correction algorithm. The obtained network provides us with a tool to understand patterns of collaborations in transportation research. The results show an increasing trend of collaboration over the studied period (1990–2015), but different journals exhibit different patterns. We characterize the structural properties of this network and compute several centrality measures to quantify the performance/impact of researchers and their collaborations in the research community. This study could serve as a tool to qualitatively and quantitatively understand scientific collaborations in transportation research.

1. Introduction

Scientific research is becoming increasingly multi-disciplinary, requiring a broad range of intellectual frameworks, skills, and techniques. With this trend, scientific collaboration has become a dominant mechanism to bring about important research advances (Katz and Martin, 1997; Wuchty et al., 2007). Collaboration is beneficial to both researchers and the progress of modern science. As a result, scientific collaborations also make teams more productive, accelerating the speed of scientific progress. Besides the rising demand for collaboration, advances in ICT (information and communications technology) also reduce the cost of communication, making international and multi-disciplinary collaborations easier than ever before.

Scientific publication is the most important proxy to access scientific advances and to understand scientific collaborations. With the trend of increasing collaboration, more research publications are now created by teams of researchers instead of single individuals (Greene, 2007). In the field of *scientometrics*, researchers have been using publication data to investigate the pattern and trend of scientific collaboration for a long time. Among different methodologies for studying collaborations, an effective approach is to investigate coauthorship networks, since joint papers are the most straightforward proxy of successful formal collaboration.

* Corresponding author at: 75 Amherst Street, E14-474D, Cambridge, MA 02142, USA.
E-mail address: sunlijun@mit.edu (L. Sun).

http://dx.doi.org/10.1016/j.tra.2017.04.011
0965-8564/

A coauthorship network can be seen as a social network built on scientific collaborations, and thus it is amenable to social network analysis (Wasserman and Faust, 1994; Barabási, 2016). With the development of complex network theories, researchers have been using network science to re-investigate the structural properties of coauthorship networks. Newman is among the first to apply complex network analysis on coauthorship networks (Newman, 2001a,b, 2004). He studied several large coauthorship networks in physics, biomedical research and computer science. In these studies, he defined different metrics to quantify the importance of vertices and edges in the network.

Subsequently, researchers in other fields started exploring other aspects of coauthorship networks. For example, For example, Barabási et al. (2002) re-investigated the structure of coauthorship network from an evolutionary perspective. investigated the structure of coauthorship networks from an evolutionary perspective. The authors considered a coauthorship network as a complex evolving network and studied the dynamics driving the evolution of the network's topology. Wagner and Leydesdorff (2005) studied the patterns of international collaboration using data from *Science Citation Index* and showed that the trend of international collaboration could be explained by the organizing principle of preferential attachment, whereby highly-connected authors tend to attract more new collaborators. To quantify authors' impact, Yan and Ding (2009) applied four centrality measures (degree, betweenness, closeness, and pagerank index) in the coauthorship network of "library and information science", and found that these measures are significantly correlated with citation counts. There are also some works focusing on particular research fields and trying to find some field-specific insights. For example, Liu et al. (2005) conducted a comprehensive analysis of the coauthorship network in "digital library". Besides network analysis, the authors also developed a new measure called *AuthorRank* to quantify authors' impact. Acedo et al. (2006) investigated coauthorship network constructed from 10 top journals in "management and organizational studies" from 1980 to 2002. Liu et al. (2005) also did a good overview of coauthorship network research. We refer the interested readers to this reference for a comprehensive literature review on coauthorship network studies.

Scientific publication in the field of transportation research has increased dramatically in the last decade, both in terms of the number of journals and articles per journal (Button, 2015). Meanwhile, the diversity of topics within transportation research has also increased, due to advances in technology and methodology. Yet, scientometric studies of the field of transportation research remain limited. Hanssen and Jørgensen (2014) studied the effect of author and article characteristics on a paper's citation counts using data from several transportation journals. The results show that articles written by authors from more than one country and articles with shorter titles tend to be more cited. Heilig and Voß (2015) systematically studied the field of public transport using historical literature data from 2009 to 2013. The authors investigated how public transport research has evolved in terms of the development of publication patterns and major topics. Using an article abstract data set extracted from 22 transportation journals, Sun and Yin (2017) detected research themes using 'topic modeling' techniques, and quantified how topic distributions evolve over time and across different journals. To the best of our knowledge, no systematic empirical research has investigated the patterns of collaboration in the transportation research community and the structure of the coauthorship network.

To fill this gap, the present study focuses on building the coauthorship network of transportation research and analyzing its structural properties using social network analysis. The remainder of this paper is organized as follows. Section 2 introduces the publication metadata used in this study and presents in detail an algorithm to correct author names. We also list several metrics in network analysis to measure the impact/importance of researchers and their collaborations. Section 3 shows the statistical properties of the obtained coauthorship network. We applied various centrality measures to identify those most influential researchers and collaborations. Section 4 concludes this study, and discusses limitations and potential future directions.

2. Methodology

In this section, we first introduce the publication data set used in building the coauthorship network. Next, we present a network-based author name correction algorithm, which can distinguish authors with identical initials and merge names referring to the same researcher but in different formats. Finally, we describe several measures quantifying the importance of authors (vertices) and their collaborations (edges) in the coauthorship network.

2.1. Data

The data set is the same as the one used for discovering research themes and trends in transportation research (Sun and Yin, 2017). The publication metadata is obtained from Web of Science (https://apps.webofknowledge.com/), consisting of articles from 22 journals in the fields of transportation research from 1990 to 2015. The 22 journals are selected as top tier in the *Science Citation Index* (SCI) under category "Transportation Science & Technology" and in the *Social Science Citation Index* (SSCI) under category "Transportation". Fig. 1 shows the number of articles of each selected journal during 1990–2015, highlighting a clear and substantial increase in volume.

The metadata includes many fields, including journal name, volume, issue, pages, authors, article title, published time, abstract, keywords and citations. As the purpose of the current study is to construct and analyze the coauthorship network, we put particular emphasis on authorship information. Single-authored articles are excluded from the data set since they do not contribute to the coauthorship network.

2.2. Author name correction

A major defect of the publication data set is that author names are often coded in different formats (initial & full name) and sometimes a single author may report his/her name using different formats in different papers. In other words, the data set contains both full names and initials of a particular author and both of them may have more than one variant. For example, we find the name of a top researcher in our data set –"Axhausen, Kay W"–was written in various formats, including "Axhausen, K", "Axhausen, KW", "Axhausen, K W" (with a space in between), "Axhausen, Kay", and "Axhausen, Kay W". Without correcting author names, for this particular researcher we will introduce five vertices, which make the underlying coauthorship network inaccurate for our analysis. These names in different formats referring to the same author should be merged in our analysis.

The other defect is that, for a common surname, the same initial in the publication data may represent different authors. For example, both "Smith, Adam" and "Smith, Andrew" have the same initial "Smith, A" and both "Williams, Carric" and "Williams, Chad" can be coded as initial "Williams, C". In the case where we observe initials "Smith, A" or "Williams, C" in a paper, we cannot tell to which author the initial refers. This case is of particular importance to those researchers with common Asian surnames, such as "Chen", "Kim", "Lee", 'Li'", "Liu", "Lee", "Park", "Wang", "Yang" and "Zhang". In this case, the same initial referring to different authors should be distinguished.

Some corrections can be made manually. However, this is infeasible, due to the large number of ambiguous cases in our data set. In scientometrics, author name disambiguation is an important research question, and various methods have been proposed such as clustering and coauthor-based methods (Han et al., 2005; Kang et al., 2009; Strotmann et al., 2009). We refer interested readers to Kang et al. (2009) for a brief review and recent contributions to this problem. In this study, we develop a new author name correction algorithm to update the publication data (Algorithm 1). The principle of the proposed algorithm is to (1) distinguish identical initials that may refer to different authors (e.g., "Smith, A") and (2) merge full names/initials that may correspond to the same author (e.g., "Axhausen, Kay" & "Axhausen, Kay W").

Algorithm 1. Author name correction.

```
Data: Collections of papers P
Result: Coauthorship network G
# initialization;
for p in P do
    for au in p.authors do
        # is initial? e.g., "Sun, L" [true]; "Sun, Lijun"[false];
        if au.name is initial then
            add a unique suffix to au.name;
            # e.g., add a number and full name changed "Sun, L" to "Sun, L_123";
            # this is to ensure the same initial in different articles are initialized as different vertices;
# update author names in P;
m ← 1;
while m > 0 do
    G ← CreateGraph (P);
    M ← MergeNames (G);
    m ← |M|;
    P ← UpdatePapers (P,M);
return G;
```

Algorithm 2. Functions used in Algorithm 1.

```
Function CreateGraph (G)
    # each vertex is a unique name (including suffix) in P;
    G ← graph connecting coauthors in all papers in P;
    return G;

Function MergeNames (G)
    U ← all author names in G;
    W ← empty list;
    M ← empty dictionary;
    foreach pair (au1, au2) from U do
        if au1.surname = au2.surname then
            if au1.firstname match au2.firstname then
                if Nei (G,au1,1) ∩ Nei (G,au2,1) ≥ Cr1 or Nei (G,au1,2) ∩ Nei (G,au2,2) ≥ Cr2 then
                    # flag registers whether au1 or au2 exist in any set in W;
                    e ← 0;
                    for w in W do
                        if au1 ∈ w or au2 ∈ w then
                            w ← w ∩ {au1, au2};
                            e ← 1;
                            break;
                    if e = 0 then W.add({au1, au2});
    for w in W do
        aut ← longer/longest name in w;
        for au ∈ w do
            if au ≠ aut then M[au] ← aut;
    return M;

Function UpdatePapers (P,M)
    # flag registers if any paper updated;
    for p in P do
        for au in p.authors do
            if au in M then
                au ← M[au];
    return P;

Function Nei (G,au,n)
    return set of n^th order neighbors of au in G;
```

In the initialization stage, we first distinguish initials in all papers. In doing so, we assign each initial with a unique suffix code (e.g., a number). For example, if we observe two "Smith, A" in two different papers, each of them will be assigned a unique number, and thus they become distinguishable. After this procedure, the collection of papers P is updated with each initial name transformed to a new unique label (e.g., "Smith, A" to "Smith, A_123"). Note that we will introduce more vertices than expected at this stage, since identical initials referring to the same author will be differentiated.

After the initialization stage, the algorithm creates a crude coauthorship network G for the purpose of detecting names to be merged. Although this network contains more vertices than expected, it still captures the right structural information from those papers with authors' full names available. Therefore, we could take advantage of this structural information to identify author names/initials that should be merged. The identification of cases to be merged is based on a name-neighbor matching principle. First, if two vertices in G have the same surname and their firstnames satisfy name matching criterion, we consider that they might be the same researcher and proceed to check their neighbor matching criterion. We define a positive neighbor matching as "two authors share at least Cr_1 first order neighbors or at least Cr_2 second order neighbors in G" (the nth order neighbors of a vertex i are vertices that are reachable from i in at most n steps). If a pair of

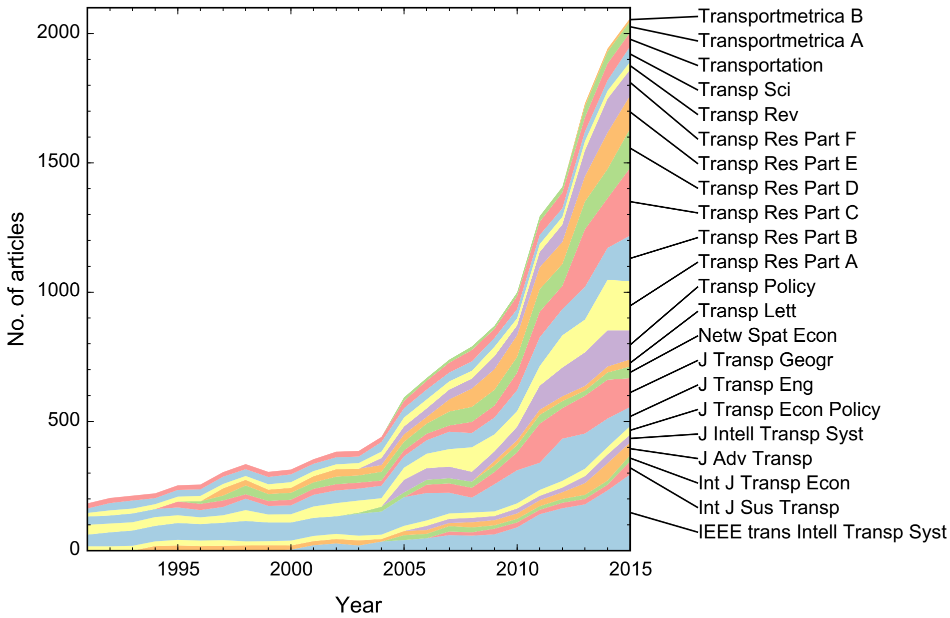

Fig. 1. Number of articles for each journal during 1990–2015.

authors meets both the name matching and the neighbor matching criteria, we consider these two names correspond to the same researcher. For better performance, we put all these possible pairs to an updating list W, in which each element is a set of names to be unified. W is initialized to be an empty list. After each positive pair, W is updated to include those unseen author names (see Algorithm 1). For example, after three positive pairs (“Sun, L_123”, “Sun, Lijun”), (“Axhausen, K_587”, “Axhausen, KW”) and (“Axhausen, KW”, “Axhausen, Kay W”), W will be updated to a list of two sets W = [{“Sun, L_123”, “Sun, Lijun”}, {“Axhausen, K_587”, “Axhausen, KW”, “Axhausen, Kay W”}].

We next build an empty dictionary M, which is prepared to further update the collection of papers P. The dictionary M is created as follows: for each element $w \in W$ and for each name $au \in w$, we pick the longer or the longest full name in w as a targeted new name au_t (if all names in w are initials, we pick the one with larger suffix number), and for all $au \in w$ and $au \neq au_t$, we assign $M[au] = au_t$. Using this dictionary, we can update each author name in M to a targeted new version.

The proposed algorithm repeats this “create graph–merge names–update papers” process. In each iteration, if there exists at least one name to be updated in M, we continue building a new coauthorship network G. The procedure stops when no further names can be updated in P. The obtained network G in this iteration is the final coauthorship network. The core of this procedure is to identify names to be merged using their neighbor structure in the coauthorship network G. The key parameters are the critical sizes of common collaborators (neighbors) Cr_1 and Cr_2. These two parameters control the levels of name detection and merging. A large value tends to merge names referring to different authors (over-detection), and a small value may overlook some cases (under-detection). In practice, one can increase/decrease the critical values Cr_1 and Cr_2 or account for an even higher order of neighbors (e.g., $\mathtt{Nei}(G, au_1, 3) \cap \mathtt{Nei}(G, au_2, 3) \geqslant Cr_3$) to find more merging cases or prevent wrong merging cases.

2.3. Network analysis

After the name correction process, we obtain the final coauthorship network $G(V, E)$ (V is the set of vertices (authors) and E is the set of edges (collaborations)) for the structural analysis.

We first introduce some measures to quantify the importance of vertices and edges in this network. The following notations are used in defining these measures: we use i, j to denote vertices (authors) and (i, j) to denote an edge connecting authors i and j; N_i represents the set of neighbors of author i; and au_p is the set of authors of paper p; $|au_p|$ is the number of authors in paper p. Note that network G here is undirected and we always have $|au_p| \geqslant 2$ since we only include papers with at least two authors in the collection.

A simple measure is degree d_i, which counts the number of connections (or neighbors) of vertex i:

$$d_i = \sum_{j \in V} a_{ij}, \tag{1}$$

where a_{ij} equals to 1 if there exists an edge connecting i and j and 0 otherwise.

For an edge (i,j), we define its "occurrence" u_{ij} as the total number of papers that authors i and j have collaborated together:

$$u_{ij} = \sum_{p \in P} \mathbb{I}\left(i \in au_p \text{ and } j \in au_p\right), \tag{2}$$

where $\mathbb{I}(e) = 1$ if condition e is true and 0 otherwise.

Note that in quantifying occurrence u, all papers have equal weight no matter how many authors appear in the paper. To justify the inflation by number of authors, a different weight parameter w was proposed in Newman (2001b):

$$w_{ij} = \sum_{p \in P} \frac{1}{|au_p| - 1} \mathbb{I}\left(i \in au_p \text{ and } j \in au_p\right), \tag{3}$$

with each collaborated paper contributing $\frac{1}{|au_p|-1}$ unit to the total weight.

We define total strength s_i of researcher i using the weight parameter as:

$$s_i = \sum_{j \in N_i} w_{ij}. \tag{4}$$

Given the formulation of w_{ij}, the total weight coming from a single paper for an author i always equals to one. In other words, we have $s_i = \sum_{p \in P} \mathbb{I}(i \in au_p)$ and it actually counts the total number of collaborated papers (total research output/productivity) of an author.

The measures above are all based on local information. We next introduce some measures that account for the global effect: betweenness (for vertices and edges) and pagerank index (for vertices). Betweenness is a shortest path-based centrality measure, quantifying the importance of a vertex in passing information along shortest paths in the network. We denote δ_{jk} as total number of shortest paths between j and k, and $\delta_{jk}(i)$ as the number of shortest paths that pass through vertex i $(i \neq j,\ i \neq k)$. Then pair (j,k) contributes an amount of $\delta_{jk}(i)/\delta_{jk}$ to the betweenness centrality of vertex i. Note that for a particular pair of vertices (j,k) $(j \neq k)$, there may exist more than one shortest path.

The total betweenness of vertex i is computed by summing the contribution from all vertex pairs in the network

$$bv_i = \sum_{i \neq j, i \neq k, j \neq k} \frac{\delta_{jk}(i)}{\delta_{jk}}. \tag{5}$$

Under the same principle, the betweenness centrality of an edge (i,j) is defined as:

$$be_{ij} = \sum_{h \neq k} \frac{\delta_{hk}[ij]}{\delta_{hk}}, \tag{6}$$

where $\delta_{hk}[ij]$ is the number of shortest paths between h and k that pass through edges (i,j).

In these two measures, we assume all edges have unit cost. Taking the collaboration strength into account, we can also compute the weighted version of betweenness bv_i^w and be_{ij}^w. The weighted cost function should be a monotone decreasing function of w_{ij}, since edges with stronger collaboration will have a shorter distance. In this study, we use $1/w_{ij}$ as weighted edge cost.

The last measure used to quantify vertex importance is the pagerank index pr_i, which is computed by performing the PageRank algorithm on the network. Pagerank index can also be considered a variant of eigenvector centrality. The principle of the algorithm is to estimate both the number and the quality of links connected to a vertex. We refer interested readers to Page et al. (1999) for the details of this algorithm. One crucial parameter in the PageRank algorithm is the damping factor d_p, which determines the probability of continuing a random walk (and $1 - d_p$ is the probability of restarting a new random walk from an arbitrary vertex). The suggested d_p is 0.85 and we use this value in our analysis. We apply the algorithm on both the binary version (pr_i, unit cost) and the weighted version (pr_i^w, using w_{ij} as edge weight) of the coauthorship network.

Apart from these network-based measures, the publication metadata also includes the number of citations of each paper, which is also a good indicator to quantify author importance. We denote the total number of citations of an author i by:

$$c_i = \sum_{p \in P} ct_p \times \mathbb{I}(i \in au_p), \tag{7}$$

where ct_p is the number of citations of paper p. It should be noted that c_i does not include citations from single-authored papers.

We summarize the measures used to quantify importance of vertice and edges in Table 1.

Table 1
List of measures used in this paper.

Vertex (author)		Edge (collaboration)	
Degree	d_i	Occurrence	u_{ij}
Strength	s_i	Weight	w_{ij}
Betweenness	bv_i	Betweenness	be_{ij}
Betweenness [weighted]	bv_i^w	Betweenness [weighted]	be_{ij}^w
Pagerank	pr_i		
Pagerank [weighted]	pr_i^w		
Citation	c_i		

Table 2
Author & collaboration statistics in each journal.

Journal	Papers	Single	Authors	Authors per paper	(max)	Collaboration	Papers per author	Citations per paper
IEEE trans Intell Transp Syst	1480	50	3790	3.52	17	8599	1.38	13.63
Int J Sus Transp	200	33	487	2.75	7	633	1.13	5.37
Int J Transp Econ	188	55	363	2.21	7	358	1.14	1.68
J Adv Transp	509	88	934	2.53	7	1315	1.38	5.18
J Intell Transp Syst	210	7	562	3.11	8	844	1.16	6.97
J Transp Econ Policy	523	205	723	1.92	10	715	1.39	11.74
J Transp Geogr	898	188	1688	2.57	18	2707	1.37	9.28
J Transp Eng	2122	267	3636	2.64	9	5922	1.54	7.68
Netw Spat Econ	312	43	620	2.71	8	951	1.36	9.08
Transp Policy	852	163	1661	2.53	10	2360	1.30	8.34
Transp Rev	629	219	1054	2.12	6	1112	1.26	11.48
Transportation	789	160	1239	2.40	8	1824	1.53	16.31
Transp Lett	144	13	330	2.85	8	480	1.24	2.33
Transp Res Part A	1610	322	2900	2.50	11	4278	1.39	17.06
Transp Res Part B	1519	309	1816	2.38	9	3313	1.99	25.22
Transp Res Part C	1314	99	2783	3.04	10	5175	1.44	13.74
Transp Res Part D	1082	146	2479	2.88	13	4081	1.26	12.18
Transp Res Part E	1066	159	1863	2.55	7	2651	1.46	14.91
Transp Res Part F	711	64	1612	3.14	14	3220	1.39	9.55
Transp Sci	797	86	1354	2.64	7	2154	1.55	27.91
Transportmetrica A	253	35	521	2.73	7	763	1.33	6.94
Transportmetrica B	36	7	98	2.92	5	127	1.07	3.47

3. Results

In this section, we first present the analysis of authors and collaborations at the journal level using the updated publication metadata with name correction. Next, we study the structural properties of the coauthorship network. Finally, the introduced measures are applied to quantify the importance of authors and their collaborations.

3.1. Journal statistics

Table 2 lists the 22 journals and their basic statistics regarding author/collaboration, including the total number of papers, the number of single-authored papers, the total number of unique authors, the average number of authors per paper (and the maximum number), the total number of collaborations, the average number of published papers per author, and average number of citations per article.

The average number of authors per paper is computed as $\bar{n} = \sum_{p\in P}|au_p|/|P|$, where $|P|$ is the total number of papers in the journal. The total number of collaborations in a paper p is $C^2_{|au_p|} = \frac{|au_p|(|au_p|-1)}{2}$ $(|au_p| \geqslant 2)$. For the whole journal, the value is $\sum_{p\in P}C^2_{|au_p|}$. The last column in Table 2 shows the average number of citations per paper for each journal. *Transportation Science*, *Transportation Research Part B: Methodological* and *Transportation Research Part A: Policy and Practice* are the top three journals with the highest number of citations per article.

As shown in Table 2, *IEEE Transactions on Intelligent Transportation Systems* has 3.52 authors per paper on average, which is the highest among all journals. The minimum average number 2.12 comes from the journal *Transport Reviews*. We take these two journals as examples to study the distribution and trend of number of authors per paper. The bar plot in Fig. 2A shows the distribution of number of authors across all papers, and the two curves show the same distribution for *IEEE Transactions on Intelligent Transportation Systems* and *Transport Reviews*, respectively. As we can see, most papers in the data set have 1–4 authors and the average number of authors from all papers in the data set is 2.68 ± 1.30 (mean $\pm$ standard deviation). However, different journals seem to have different signatures in terms of this distribution. For example, about 35% of papers in *Transport Reviews* are single-authored, while the fraction is less than 5% in *IEEE Transactions on Intelligent Transportation Systems*. In the whole data set, there exist two extreme observations with more than 15 authors: "Cooperative adaptive

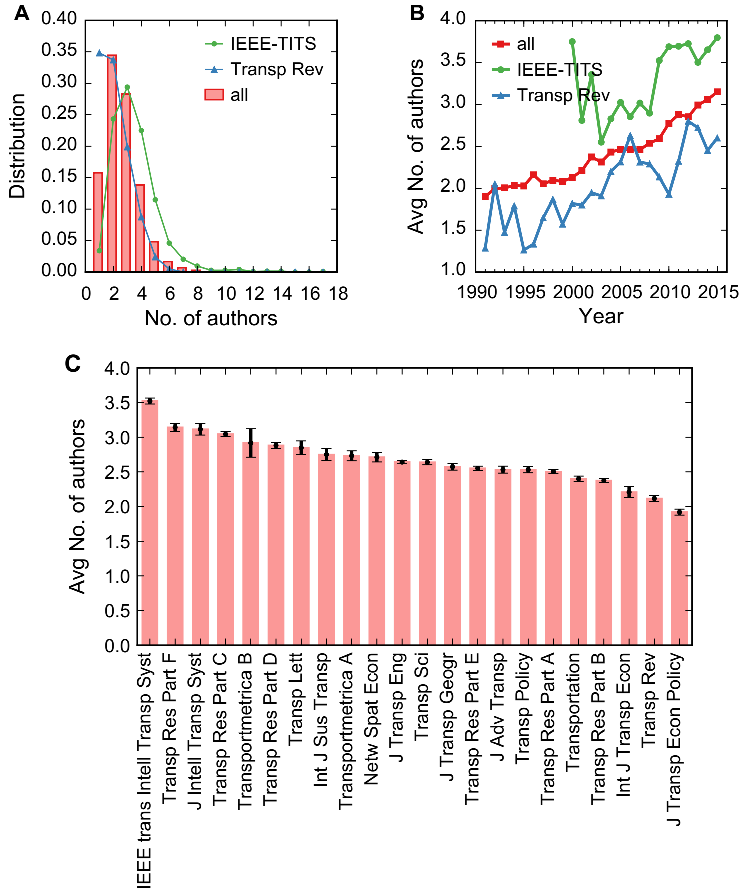

Fig. 2. Basic statistics in publication metadata. (A) distribution of number of authors per paper; (B) the temporal variation of average number of authors per paper; and (C) average number of authors per paper for each journal. The error bar shows the standard error of the mean.

cruise control implementation of team mekar at the grand cooperative driving challenge." from *IEEE Transactions on Intelligent Transportation Systems* with 17 authors and "Visions for a walking and cycling focussed urban transport system" from *Journal of Transport Geography* with 18 authors.

Fig. 2B plots how the average number of authors per paper varies with time. We observe a clear increasing trend from 1.8 to 3.2 during 1990–2015 (the red[1] curve), suggesting that collaborations are becoming more and more common over time. *Transport Reviews* follows the same trend but with a shift towards fewer authors. *IEEE Transactions on Intelligent Transportation Systems* on the other hand essentially has a larger number of authors than the average across all journals. In fact, trend of increasing collaboration is observed in each of the 22 journals as well.

This natural trend of increased scientific collaboration has often been attributed to two factors. On one hand, making larger technical contributions requires combining multiple methodologies and expertise from different fields, something costly for small teams of authors to achieve. On the other hand, the process of collaboration itself, and in particular international collaboration, becomes less costly with the help of ICT (Wagner and Leydesdorff, 2005; Adams, 2012). The effect of these factors is consistent with our the observed the number of authors per paper by journal–see Fig. 2C. Essentially, we found that journals in methodology, planning, economics and policy — such as *Journal of Transport, Economics and Policy*, *Transport Reviews*, *International Journal of Transport Economics*, *Transportation Research Part B: Methodological*, *Transportation*, and *Transportation Research Part A: Policy and Practice*—tend to involve fewer authors per paper, compared to journals in technology such as *IEEE Transactions on Intelligent Transportation Systems* and *Transportation Research Part C: Emerging Technology*. Presumably, this is driven by the need to bring a broader range of technical expertise to make technological contributions.

3.2. Structural properties of the coauthorship network

We build the coauthorship network by applying Algorithm 1 with $Cr_1 = 3$ and $Cr_2 = 5$. The constructed coauthorship network contains 22,920 vertices and 43,163 edges. Fig. 3 visualizes its largest connected component (LCC), which consists of

[1] For interpretation of color in Fig. 2, the reader is referred to the web version of this article.

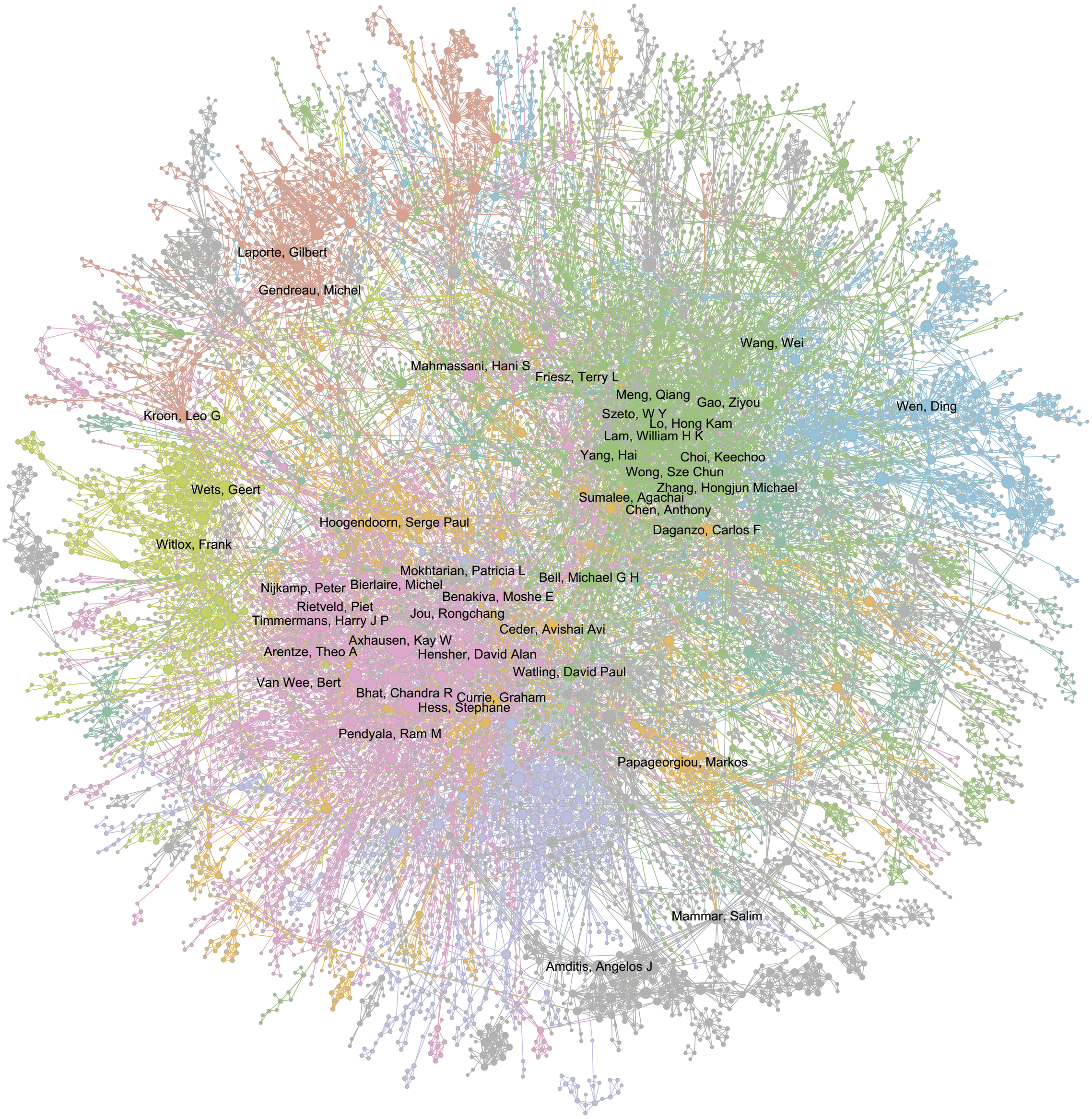

Fig. 3. The largest connected component in the coauthorship network. The names of those authors with $d_i \geqslant 40$ are shown.

12,273 vertices and 29,661 edges. For better visualization, we use different colors to indicate the community structure identified by the Louvain method (Blondel et al., 2008) and label names of those authors with $d_i \geqslant 40$. We also show in Fig. 4 the connections among the top 200 authors with the highest degrees as a core structure of the coauthorship network. Those vertices not presented in the LCC in general form small cliques. These cliques are mostly constructed by those authors who do not collaborate with the general community.

We next show some statistical properties about this network. Panels (A), (B) and (C) in Fig. 5 show the distributions of degree d, strength s and citation c, respectively. All these three distributions have heavy tails, which are well characterized by a power-law. We plot the best power-law fit $P(x) \sim x^{\beta}$ for these distributions as green lines in these panels. These power-law tails indicate great heterogeneity and inequality in the coauthorship networks: most of the authors have a small centrality value, while there still exist some author with a very high degree, strength and number of citations.

For degree d, we observe a "fat tail" with $\beta = -2.89$. The top three authors with highest degrees are "Axhausen, Kay W" (104 coauthors), "Lam, William H K" (99 coauthors) and "Yang, Hai" (98 coauthors). In terms of strength s (or productivity, the total number of papers from an author), this heavy tail ($\beta = -3.52$) suggests that many researchers only publish a few papers, while a small number of authors do publish many papers. In fact, 71.5% of the authors only published one paper, while the maximum value is 139 from "Hensher, David Alan". Fig. 5C shows a power-law tail with $\beta = -2.58$ of number

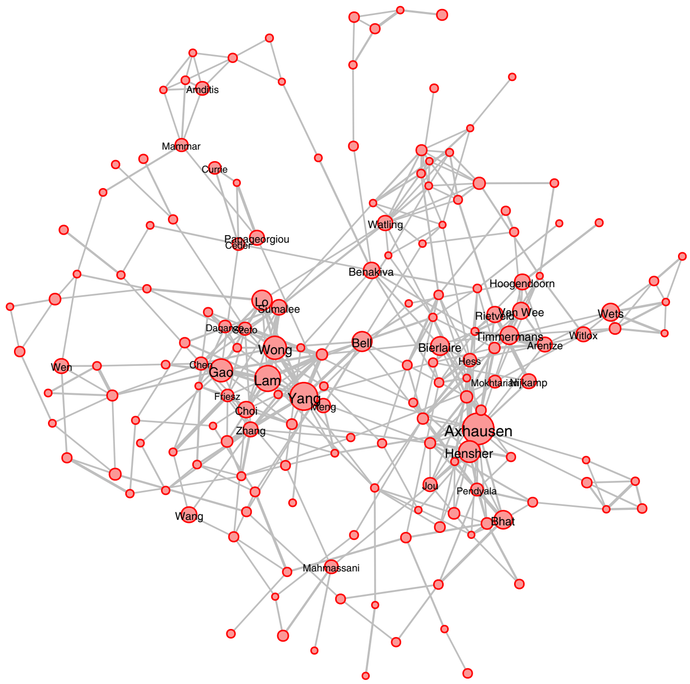

Fig. 4. Coauthorship network built from 200 authors with highest degrees (for simplicity, only the largest component is shown here; labels show the surname of authors with $d(i) \geqslant 40$).

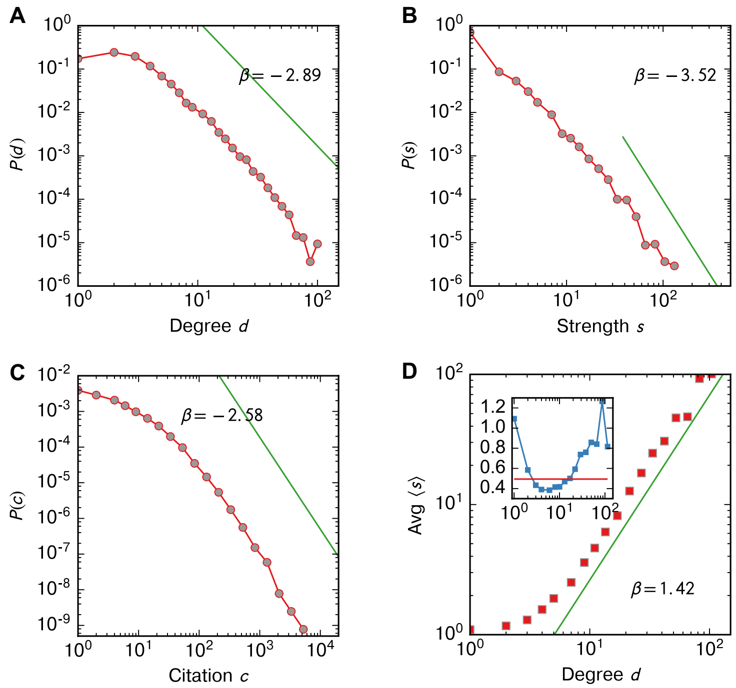

Fig. 5. Network statistics: (A) degree distribution $P(d)$; (B) strength distribution $P(s)$; (C) citation distribution $P(c)$. In panel (A), (B), and (C), the dotted line shows the distribution and the green line with annotation shows a power-law fit with estimated exponent. (D) Average strength $\langle s \rangle$ as a function of degree d of vertices. The inset shows the variation of $\langle s/d \rangle$ as a function of d.

of citations. In this network, we find that 13,900 authors received less than ten citations and only 26 of them received more than 1000 citations (without accounting for single-authored papers). When single-authored papers are included, the number of authors with more than 1000 citations becomes 30. This also demonstrates the great inequality/heterogeneity in research

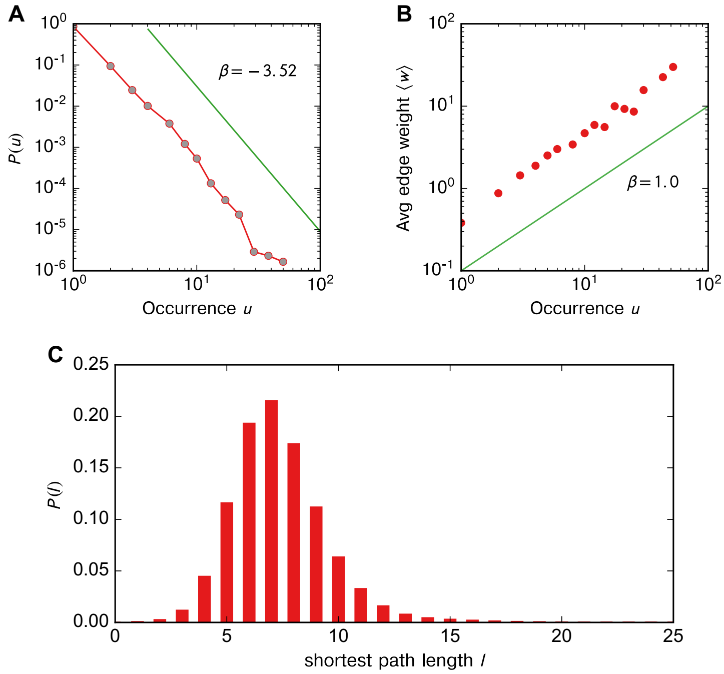

Fig. 6. Network statistics: (A) edge occurrence distribution $P(u)$; (B) average edge weight $\langle w \rangle$ as a function of occurrence u of edges. (C) shortest path length distribution in the coauthorship network.

performance. These patterns were also observed in other fields such as physics, biomedical research, and computer science (Newman, 2001a).

Panel (D) shows the average strength $\langle s \rangle$ as a function of degree d. In the case where edge weights are randomly distributed, one expects $\langle s \rangle$ is in proportion to vertex degree d. However, in this network, we observe that $\langle s \rangle \sim d^{\beta}$ for $d \geqslant 10$ with $\beta \approx 1.42$, suggesting that total output of authors increases faster than their degrees. It also implies that edges from highly connected authors are stronger than other edges.

To further investigate this, we show $\langle s/d \rangle$ as a function of d in the inset of Panel (D). We observe an interesting non-linear relation, in which $\langle s/d \rangle$ first decreases for $1 \leqslant d \leqslant 5$ and then goes up for $d \geqslant 6$ (the blue squares). While in a randomly distributed case, we will observe a constant $\langle w_{ij} \rangle = 0.51$. This may characterize the dynamics of how researchers establish themselves in different stages of their careers. When an author only has a few coauthors ($d \leqslant 2$), the connection he/she establish with coauthors tends to be strong and dedicated. This may refer to the junior-senior relation (e.g., students and supervisors) or small but isolated cliques mentioned before. For those authors with $3 \leqslant d < 20$, they are prone to build more connections with the development of career, and thus $\langle s/d \rangle$ becomes lower than a randomized case. The last stage for $d \geqslant 20$ may represent those established researchers who have stable and active collaborations with their coauthors. In this stage, the increase in collaboration with existing collaborators is faster than creating new connections.

In terms of edges, Fig. 6A shows the distribution of edge occurrence u. This is also characterized by a power-law with $\beta = -3.52$. From all edges, 85.9% have $u = 1$ and only 1.2% have $u \geqslant 5$. Fig. 6B shows the relationship between average edge weight $\langle w \rangle$ and occurrence u. This is also almost captured by a linear relation $\langle w \rangle \sim u$, suggesting that the number of other authors of papers between a pair of authors (i, j) seems to be invariant of edge occurrence u_{ij}.

Fig. 6C shows the shortest path length distribution $P(l)$. The average (connected) path length in this network is $\langle l \rangle = 7.31$. Since the network is not connected as a full component, we quantify the maximum (reachable) distance in the largest connected component, which is 28. The values of average path length essentially suggest the small world property of the coauthorship network: on average two authors are connected in only about 7 hops.

3.3. Identify influential researchers

In this section we apply the introduced measures in Section 2 to quantify researchers' importance in the coauthorship network. The first three measures are degree d, strength s and citation c presented in Fig. 5. Those influential researchers are located at the tails of these three centrality distributions. The other two indicators are betweenness centrality bv, and pagerank index pr. Tables 3 and 4 list those top 30 researchers identified using different measures.

Degree d simply counts the number of edges (i.e., the number of unique collaborators) of an author. The degree of these top researchers ranges from 45 to 104, while the average number is 3.76 at the level of the whole network. When measuring

Table 3
Ranking of authors based on coauthorship degree, coauthorship strength, number of citations, and total number of citations with single-authored papers included.

	Degree d		Strength s		Citation c		Citation [with single-authored] c^s	
1	Axhausen, Kay W	104	Hensher, David Alan	139	Yang, Hai	4281	Yang, Hai	4643
2	Lam, William H K	99	Yang, Hai	138	Hensher, David Alan	3342	Daganzo, Carlos F	4130
3	Yang, Hai	98	Lam, William H K	114	Mokhtarian, Patricia L	2769	Hensher, David Alan	3937
4	Wong, Sze Chun	88	Wong, Sze Chun	102	Lam, William H K	2274	Bhat, Chandra R	3390
5	Gao, Ziyou	79	Timmermans, Harry J P	87	Wong, Sze Chun	1985	Mokhtarian, Patricia L	2915
6	Hensher, David Alan	78	Meng, Qiang	82	Bhat, Chandra R	1960	Lam, William H K	2274
7	Lo, Hong Kam	73	Lo, Hong Kam	74	Lo, Hong Kam	1718	Wong, Sze Chun	2170
8	Bell, Michael G H	66	Bhat, Chandra R	73	Laporte, Gilbert	1531	Bell, Michael G H	2011
9	Wang, Wei	64	Rietveld, Piet	70	Rietveld, Piet	1503	Lo, Hong Kam	1899
10	Timmermans, Harry J P	63	Arentze, Theo A	60	Meng, Qiang	1489	Laporte, Gilbert	1675
11	Bierlaire, Michel	61	Zhang, Hongjun Michael	58	Daganzo, Carlos F	1469	Rietveld, Piet	1626
12	Wets, Geert	61	Gao, Ziyou	56	Papageorgiou, Markos	1466	Papageorgiou, Markos	1595
13	Bhat, Chandra R	60	Rose, John M	55	Bell, Michael G H	1438	Meng, Qiang	1489
14	Choi, Keechoo	58	Huang, Haijun	54	Gendreau, Michel	1333	Handy, Susan L	1458
15	Rietveld, Piet	57	Mokhtarian, Patricia L	53	Handy, Susan L	1293	Zhang, Hongjun Michael	1456
16	Laporte, Gilbert	56	Bell, Michael G H	52	Mahmassani, Hani S	1254	Gendreau, Michel	1333
17	Benakiva, Moshe E	55	Chen, Anthony	51	Timmermans, Harry J P	1236	Cervero, Robert	1289
18	Van Wee, Bert	55	Axhausen, Kay W	50	Benakiva, Moshe E	1190	Mahmassani, Hani S	1254
19	Zhang, Hongjun Michael	52	Sumalee, Agachai	49	Kitamura, Ryuichi	1190	Timmermans, Harry J P	1236
20	Hoogendoorn, Serge Paul	51	Benakiva, Moshe E	48	Greene, William H	1133	Kitamura, Ryuichi	1209
21	Sumalee, Agachai	50	Papageorgiou, Markos	48	Axhausen, Kay W	1127	Benakiva, Moshe E	1190
22	Mahmassani, Hani S	49	Van Wee, Bert	46	Chen, Anthony	1115	Huang, Haijun	1170
23	Meng, Qiang	49	Mahmassani, Hani S	46	Huang, Haijun	1114	Savelsbergh, Martin W P	1154
24	Watling, David Paul	49	Yin, Yafeng	44	Trivedi, Mohan Manubhai	1102	Greene, William H	1133
25	Arentze, Theo A	48	Jaradiaz, Sergio R	44	Rose, John M	1053	Axhausen, Kay W	1127
26	Nijkamp, Peter	48	Fwa, Tien Fang	44	Mannering, Fred L	1035	Chen, Anthony	1115
27	Papageorgiou, Markos	48	Daganzo, Carlos F	43	Cao, Xinyu Jason	979	Trivedi, Mohan Manubhai	1102
28	Wen, Ding	48	Rakha, Hesham A	42	Koppelman, Frank S	948	Mannering, Fred L	1056
29	Witlox, Frank	48	Nijkamp, Peter	41	Cervero, Robert	944	Rose, John M	1053
30	Abdelaty, Mohamed A	45	Zhang, Am	41	Kockelman, Kara M	918	Cao, Xinyu Jason	1003

Table 4
Ranking of authors based on betweenness, weighted betweenness, pagerank index and weighted pagerank index.

	Betweenness bv ($\times 10^6$)		Betweenness [weighted] bv^w ($\times 10^6$)		Pagerank pr ($\times 10^{-4}$)		Pagerank [weighted] pr^w ($\times 10^{-4}$)	
1	Axhausen, Kay W	6.28	Yang, Hai	22.22	Axhausen, Kay W	7.48	Yang, Hai	12.27
2	Yang, Hai	5.75	Hensher, David Alan	13.11	Lam, William H K	6.65	Hensher, David Alan	12.11
3	Benakiva, Moshe E	3.56	Wong, Sze Chun	11.93	Yang, Hai	6.63	Lam, William H K	10.41
4	Bierlaire, Michel	3.49	Lo, Hong Kam	10.45	Hensher, David Alan	6.29	Wong, Sze Chun	9.24
5	Mahmassani, Hani S	3.40	Meng, Qiang	8.80	Wong, Sze Chun	5.70	Bhat, Chandra R	8.62
6	Gao, Ziyou	3.31	Bell, Michael G H	8.19	Lo, Hong Kam	5.33	Rietveld, Piet	8.29
7	Lam, William H K	3.30	Zhang, Hongjun Michael	6.94	Gao, Ziyou	5.04	Timmermans, Harry J P	8.22
8	Hensher, David Alan	3.27	Rose, John M	6.52	Bierlaire, Michel	4.93	Lo, Hong Kam	7.48
9	Lo, Hong Kam	3.25	Recker, Wilfred W	6.12	Bell, Michael G H	4.84	Meng, Qiang	7.42
10	Bell, Michael G H	2.66	Daganzo, Carlos F	6.10	Rietveld, Piet	4.78	Axhausen, Kay W	6.99
11	Szeto, W Y	2.63	Lam, William H K	6.09	Bhat, Chandra R	4.78	Mokhtarian, Patricia L	6.37
12	Zhang, Hongjun Michael	2.49	Golob, Thomas F	5.75	Laporte, Gilbert	4.69	Gao, Ziyou	6.29
13	Peeta, Srinivas	2.48	Fujii, Satoshi	5.74	Rakha, Hesham A	4.47	Mahmassani, Hani S	6.23
14	Kitamura, Ryuichi	2.18	Hess, Stephane	5.48	Timmermans, Harry J P	4.43	Bell, Michael G H	6.19
15	Daganzo, Carlos F	2.06	Schmoecker, Jandirk	4.83	Wang, Wei	4.40	Benakiva, Moshe E	5.94
16	Khattak, Asad J	2.00	Ettema, Dick F	4.77	Mahmassani, Hani S	4.38	Laporte, Gilbert	5.86
17	Geroliminis, Nikolas	2.00	Zhang, Xiaoning	4.67	Van Wee, Bert	4.28	Rakha, Hesham A	5.76
18	Meng, Qiang	1.96	Jaradiaz, Sergio R	4.67	Benakiva, Moshe E	4.10	Zhang, Hongjun Michael	5.69
19	Karlaftis, Matthew G	1.94	Timmermans, Harry J P	4.51	Choi, Keechoo	4.06	Arentze, Theo A	5.54
20	Van Wee, Bert	1.92	Huang, Haijun	4.50	Kockelman, Kara M	4.01	Kockelman, Kara M	5.25
21	Madanat, Samer M	1.91	Lin, Weihua	4.46	Nijkamp, Peter	3.95	Bierlaire, Michel	5.25
22	Polak, John W	1.89	Garling, Tommy	3.92	Mokhtarian, Patricia L	3.87	Daganzo, Carlos F	5.21
23	Lee, Derhorng	1.84	Mahmassani, Hani S	3.83	Jou, Rongchang	3.78	Van Wee, Bert	5.17
24	Watling, David Paul	1.82	Axhausen, Kay W	3.83	Daganzo, Carlos F	3.71	Nijkamp, Peter	5.14
25	Chen, Anthony	1.82	Holguinveras, Jose	3.81	Ceder, Avishai Avi	3.63	Papageorgiou, Markos	4.98
26	Speranza, Maria Grazia	1.79	Kitamura, Ryuichi	3.73	Zhang, Hongjun Michael	3.57	Jaradiaz, Sergio R	4.92
27	Wong, Sze Chun	1.79	Tirachini, Alejandro	3.55	Hoogendoorn, Serge Paul	3.54	Peeta, Srinivas	4.90
28	Washington, Simon P	1.78	Chorus, Caspar G	3.52	Gendreau, Michel	3.53	Levinson, David M	4.83
29	Mannering, Fred L	1.66	Benakiva, Moshe E	3.46	Meng, Qiang	3.53	Fwa, Tien Fang	4.77
30	Hoogendoorn, Serge Paul	1.59	Jou, Rongchang	3.37	Fwa, Tien Fang	3.48	Huang, Haijun	4.70

Table 5
Kendall's τ between different centrality measures.

	d	s	c	c^s	bv	bv^w	pr	pr^w
Degree	1.000	0.438	0.229	0.169	0.511	0.469	0.483	0.314
Strength	0.438	1.000	0.430	0.362	0.733	0.680	0.402	0.526
Citation	0.229	0.430	1.000	0.938	0.367	0.346	0.205	0.246
Citation [s]	0.169	0.362	0.938	1.000	0.358	0.340	0.154	0.197
Betweenness	0.511	0.733	0.367	0.358	1.000	0.890	0.462	0.475
Betweenness [w]	0.469	0.680	0.346	0.340	0.890	1.000	0.459	0.474
Pagerank	0.483	0.402	0.205	0.154	0.462	0.459	1.000	0.720
Pagerank [w]	0.314	0.526	0.246	0.197	0.475	0.474	0.720	1.000

strength s, we find that the order change substantially compared with that of degree d. For example, "Hensher, David Alan" move from the 6th to the 1st and "Timmermans, Harry J P" rise from the 10th to the 5th. These rising authors have stronger and more dedicated collaborations with their coauthors.

The number of citations ranges from 918 to 4281 for the top 30 authors (Table 3). Note that in counting citations we only include papers with more than one author to keep it consistent with those network-based measures, so it lowers the weight of those authors publishing many single-authored papers. To show the difference, we also measure the total citation c^s with single-authored papers included. We identify a special case here —"Daganzo, Carlos F", whose citation number rise substantially from 1469 to 4130. This is because he published many single-authored influential papers. Besides "Daganzo, Carlos F", the citation index is also boosted for some other authors, such as "Bhat, Chandra R" and "Bell, Michael G H".

Essentially, degree d and strength s quantify the importance of a vertex (an author) using local structural property (only first order neighbors). We next introduce the result obtained from betweenness centrality and pagerank index, which can reflect their global representations by quantifying higher order interactions using distance or random-walk-based approaches. Table 4 shows the top 30 authors identified by betweenness bv, weighted betweenness bv^w, pagerank pr and weighted pagerank pr^w. From the structural perspective, a vertex with a high betweenness tends to serve as bridges connecting different parts of the network. Correspondingly, those authors with high br is prone to establish diverse collaborations with different communities. When edge weight is considered, we find that authors with more strong collaboration become more central in terms of bv^w. This is because edge cost becomes $1/w_{ij}$ and the distance between two authors with strong collaboration will drop substantially. For example, the distance between "Arentze, Theo A" and "Timmermans, Harry J P" in this case is almost zero ($1/29.91 = 0.03$). This cost function is very sensitive to those strong collaboration, and thus bv^w may considerably promote researchers who do not have a high degree but with strong collaborations with authors in different communities. This also implies that these authors may have good interdisciplinary research experiences. Using this measure, we identify some important authors that are not among the top in terms of other measures; for example, "Recker, Wilfred W", "Golob, Thomas F", "Fujii, Satoshi", "Hess, Setephane" "Schmoecker, Jandirk", "Ettema, Dick F", "Zhang, Xiaoning", "Jaradiaz, Sergio R", etc.

A vertex's pagerank index is determined by the both the number and importance of their neighbors. In this sense, pagerank index is different from degree, strength, and betweenness, as it can capture the inherited and transferred/flowed status of vertices. More generally, we may consider pagerank a natural ranking built by authors themselves. The results from pagerank pr and weighted pagerank pr^w seem to be consistent and meaningful. These identified researchers are all well-known in transportation research, and many of them serve as editors/editorial board members in those selected journals.

In principle, the results from all centrality measures are consistent with our intuition. However, each measure places emphasis on different features. An interesting point that should be noted is that these influential vertices not only include those who are specialized in transportation, but also some renowned researchers majored in other fields who publish in transportation journals occasionally. For example, "Gendreau, Michel", "Laporte, Gilbert" and "Savelsbergh, Martin W P" from the general field of operations research and industrial engineering; "Rieveld, Piet" who works in the field of transport and urban economics; and "Trivedi, Mohan Manubhai" in computer vision. This is also true vice versa, some authors in transportation field also publish papers in operations research or other non-transportation journals. In order to compare the results from different measures, we quantify their consistency (strength of association) by computing Kendall's τ rank correlation coefficient for every pair of measures. Table 5 shows the final pairwise correlation matrix. The highest correlation 0.938 is found between citations c and overall citations c^s with single-authored paper included. This is not surprising as the difference between them is the citations of those single-authored papers. The correlation between bv and bv^w is 0.890, which is second highest value. The following pair is between strength s and betweenness bv, with $\tau = 0.733$. This high correlation indicates that the local indicator s can characterize to a certain degree the global distance-based measure bv.

3.4. Identify important collaborations

After identifying influential researchers, we focus on measuring edges in the coauthorship network. We compute occurrence u_{ij}, weight w_{ij}, betweenness be_{ij} and its weighted version be_{ij}^w, introduced in Section 2 on the network.

Table 6
Ranking of collaborations (edges) based on occurrence, weight and edge betweenness.

	Occurrence u		Weight w		Betweenness be ($\times 10^5$)		Betweenness [weighted] be^w ($\times 10^5$)	
1	Arentze, TA & Timmermans, HJP	54	Arentze, TA & Timmermans, HJP	29.91	Bierlaire, M & Speranza, MG	16.92	Wong, SC & Yang, H	89.02
2	Hensher, DA & Rose, JM	42	Hensher, DA & Rose, JM	22.50	Axhausen, KW & Geroliminis, N	10.41	Meng, Q & Yang, H	82.13
3	Meng, Q & Wang, S	27	Meng, Q & Wang, S	15.67	Benakiva, ME & Waitz, IA	10.08	Bell, MGH & Yang, H	79.26
4	Lam, WHK & Li, Z	23	Currie, G & Delbosc, A	15.17	Kitamura, R & Yang, H	9.49	Lo, HK & Wong, SC	64.84
5	Wong, SC & Yang, H	21	Hensher, DA & Li, Z	13.00	Li, H & Parent, M	8.85	Hensher, DA & Rose, JM	58.66
6	Lam, WHK & Sumalee, A	20	Greene, WH & Hensher, DA	11.03	Barth, MJ & Du, J	8.69	Golob, TF & Hensher, DA	53.29
7	Nijkamp, P & Rietveld, P	20	Tong, CO & Wong, SC	10.50	Mahmassani, HS & Murraytuite, P	8.58	Golob, TF & Recker, WW	51.83
8	Papageorgiou, M & Papamichail, I	20	Meng, Q & Yang, H	10.33	Axhausen, KW & Lee, D	8.58	Fujii, S & Schmoecker, J	47.12
9	Hensher, DA & Li, Z	19	Huang, H & Yang, H	10.00	Axhausen, KW & Bierlaire, M	7.95	Yang, H & Zhang, X	45.29
10	Currie, G & Delbosc, A	18	Lo, HK & Szeto, WY	10.00	Joshi, S & Rathi, A	7.81	Bell, MGH & Schmoecker, J	42.92
11	Dellolio, L & Ibeas, A	18	Liu, Z & Meng, Q	9.67	Dullaert, W & Szeto, WY	7.77	Lin, W & Lo, HK	42.18
12	Lam, WHK & Wong, SC	18	Wong, SC & Yang, H	9.58	Khattak, AJ & Polak, JW	7.74	Zhang, HM & Zhang, X	41.87
13	Huang, H & Yang, H	17	Huang, H & Lam, WHK	9.58	Lo, HK & Rathi, A	7.74	Hess, S & Rose, JM	41.59
14	Meng, Q & Yang, H	17	Hensher, DA & Puckett, SM	9.17	Hensher, DA & Mannering, FL	7.72	Huang, H & Yang, H	38.86
15	Tong, CO & Wong, SC	17	Easa, SM & Hassan, Y	9.17	Wirasinghe, SC & Yang, H	7.44	Fujii, S & Garling, T	35.69
16	Greene, WH & Hensher, DA	16	Papageorgiou, M & Papamichail, I	8.90	Mahmassani, HS & Peeta, S	7.43	Hensher, DA & Tirachini, A	35.53
17	Liu, Z & Meng, Q	16	Lam, WHK & Li, Z	8.58	Barth, MJ & Trivedi, MM	7.36	Huang, H & Lam, WHK	35.37
18	Friesz, TL & Yao, T	15	Meng, Q & Weng, J	8.50	Ceder, AA & Gao, Z	7.18	Gelareh, S & Meng, Q	33.31
19	Bhat, CR & Pendyala, RM	14	Dresner, ME & Windle, RJ	8.33	Benakiva, ME & Madanat, SM	6.87	Recker, WW & Zhang, HM	32.41
20	Dresner, ME & Windle, RJ	14	Nijkamp, P & Rietveld, P	8.33	Engstrom, J & Victor, T	6.86	Jaradiaz, SR & Tirachini, A	32.00
21	Friman, M & Garling, T	14	Golob, TF & Regan, AC	8.00	Fu, X & Lam, WHK	6.86	Gelareh, S & Pisinger, D	30.84
22	Fujii, S & Garling, T	14	Qian, ZS & Zhang, HM	8.00	Lin, W & Lo, HK	6.82	Daganzo, CF & Lin, W	30.57
23	Janssens, D & Wets, G	14	Lam, WHK & Tam, ML	7.08	Dullaert, W & Gendreau, M	6.73	Ettema, DF & Timmermans, HJP	29.57
24	Lam, WHK & Tam, ML	13	De Palma, A & Lindsey, R	7.00	Benakiva, ME & Bierlaire, M	6.56	Lam, WHK & Sumalee, A	28.77
25	Lawphongpanich, S & Yin, Y	13	Zhang, A & Zhang, Y	7.00	Andersson, H & Meng, Q	6.51	Bliemer, MCJ & Rose, JM	28.37
26	Li, Z & Wong, SC	13	Dellolio, L & Ibeas, A	6.83	Axhausen, KW & Scott, DM	6.41	Holguinveras, J & Jaradiaz, SR	27.92
27	Cordeau, J & Laporte, G	12	Lawphongpanich, S & Yin, Y	6.83	Murphy, E & Zhou, J	6.35	Ettema, DF & Garling, T	26.49
28	Easa, SM & Hassan, Y	12	Buehler, R & Pucher, J	6.50	Dozza, M & Victor, T	6.29	Hensher, DA & Jou, R	26.21
29	Handy, SL & Mokhtarian, PL	12	Cao, XJ & Mokhtarian, PL	6.50	Friesz, TL & Peeta, S	6.23	Bell, MGH & Quddus, MA	25.49
30	Hensher, DA & Puckett, SM	12	Carey, M & Ge, YE	6.50	Murphy, E & Usher, J	6.23	Hess, S & Polak, JW	25.23

Table 6 shows the top 30 edges (collaboration between paired authors) in terms of the four measures. Occurrence simply quantifies the number of papers from a pair of authors. Weight w_{ij} can be considered a balanced version in which authors' contribution is rescaled by the number of authors in each paper. Both measures only capture the papers on the particular edge of interest, and in this sense they can be regarded as local indicators. A couple of strongly connected pairs can be identified using these two measures. For example, (''Arentze, Theo A" and ''Timmermans, Harry J P"), (''Hensher, David Alan" and ''Rose, John M"), (''Meng, Qiang" and ''Wang, Shuaian"), (''Currie, Gramham" and ''Delbosc, Alexa"), (''Hensher, David Alan" and ''Li, Zheng"), (''Meng, Qiang" and ''Yang, Hai"), (''Huang, Haijun" and ''Yang, Hai"), and (''Papageorgiou, Markos" and ''Papamichail, Ioannis").

The shortest path-based measure *be* and be^w capture how well a particular edge is placed in the coauthorship network in terms of connecting others together. The collaborations with high betweenness are not necessarily strong regarding occurrence and weight, but they tend to be in a more central position in shaping the structure of the network. In this sense, collaborations connecting different communities appear to have high betweenness values, as they are more likely to be on the shortest paths.

4. Conclusion and discussion

We have studied the coauthorship network structure using scientific publications in the field of transportation research. In this network, a vertex represents an author and an edge exists if two authors have collaborated in at least one paper. The publication metadata (Sun and Yin, 2017) covers 22 top tier journals in transportation research from 1990 to 2015.

To solve the author name disambiguation problem, we have developed an efficient network-based algorithm, which follows a ''create coauthorship network–merge author names–update papers" procedure. Since we only distinguish initials in the pre-processing stage, the algorithm still cannot distinguish two authors with identical full names. The other limitation is that we cannot handle author name changes due to marriage or adoption. These rare cases might be solved by further integrating authors' affiliation, although it is beyond the scope of our study. One potential solution is to build the network with a higher resolution (e.g., at the level of journals), and then follow the same procedure with full names distinguished as well. Overall, this developed algorithm works well in our study and special attentions should be paid when apply it on large research fields with full name disambiguation. Despite working on the correction algorithm, we think a potential solution is to encourage all authors to participate in author identification projects such as ORCID (http://orcid.org/) and ResearcherID (http://www.researcherid.com/). And then authors can be distinguished using their unique IDs instead of names in various formats.

From the publication data we find that scientific collaboration has been a natural and increasing trend. The overall network exhibits the ''small-world" and ''scale-free" properties. To better quantify the structure of the coauthorship network, we have computed a number of network-based centrality measures for both vertices and edges, including vertex degree, vertex strength, vertex betweenness, vertex pagerank index, edge occurrence, edge weight, and edge betweenness. To take into account collaboration strength, we also use the weighted variations of betweenness and pagerank index. The publication metadata also provides us with the number of citations of each paper. We use the total number of citations of each author as an indicator of his/her research performance. We applied these centrality metrics to quantify authors' performance/impact and identified those most prominent researchers in our data set. Those influential researchers are consistent with our intuition.

In summary, our study has provided a tool to understand scientific collaborations and the patterns of coauthorship in transportation research. From this network, we can develop a series of network-based metrics, which can be used as alternatives to evaluate researchers' performance and impact. As we have discussed, different metrics place emphasis on different features and aspects of the network structure. Therefore, one needs to be careful in selecting the right metric for a particular type of evaluation. A good indicator could be a combination of different status metric in the coauthorship network, and also include the number of citations since many researchers prefer single-authored papers. Taken together, an integration of network based centrality measures and citation-based measures could be more advantageous. A possible future direction is to integrate temporal data and quantify the evolution process of the coauthorship network, such as presented in Börner et al. (2004). This could reveal how researchers' importance/impact varies with time and different stages in their career.

Acknowledgement

This work was supported by the Center for Complex Engineering Systems (CCES) at King Abdulaziz City for Science and Technology (KACST) and the Massachusetts Institute of Technology (MIT). We would like to thank two anonymous referees for their comments and suggestions.

References

Acedo, F.J., Barroso, C., Casanueva, C., Galán, J.L., 2006. Co-authorship in management and organizational studies: an empirical and network analysis*. J. Manage. Stud. 43 (5), 957–983.
Adams, J., 2012. Collaborations: the rise of research networks. Nature 490 (7420), 335–336.
Barabási, A.-L., 2016. Network Science. Cambridge University Press.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T., 2002. Evolution of the social network of scientific collaborations. Phys. A: Stat. Mech. Appl. 311 (3), 590–614.
Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. 2008 (10), P10008.
Börner, K., Maru, J.T., Goldstone, R.L., 2004. The simultaneous evolution of author and paper networks. Proc. Natl. Acad. Sci. 101 (suppl 1), 5266–5273.
Button, K., 2015. Publishing transport research: are we learning much of use? Transp. Rev. 35 (5), 555–558.
Greene, M., 2007. The demise of the lone author. Nature 450 (7173) (1165–1165).
Han, H., Zha, H., Giles, C.L., 2005. Name disambiguation in author citations using a k-way spectral clustering method. In: JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. IEEE, pp. 334–343.
Hanssen, T.-E.S., Jørgensen, F., 2014. Citation counts in transportation research. Eur. Transp. Res. Rev. 6 (2), 205–212.
Heilig, L., Voß, S., 2015. A scientometric analysis of public transport research. J. Public Transport. 18 (2), 8.
Kang, I.-S., Na, S.-H., Lee, S., Jung, H., Kim, P., Sung, W.-K., Lee, J.-H., 2009. On co-authorship for author disambiguation. Inform. Process. Manage. 45 (1), 84–97.
Katz, J.S., Martin, B.R., 1997. What is research collaboration? Res. policy 26 (1), 1–18.
Liu, X., Bollen, J., Nelson, M.L., Van de Sompel, H., 2005. Co-authorship networks in the digital library research community. Inform. Process. Manage. 41 (6), 1462–1480.
Newman, M.E., 2001a. Scientific collaboration networks. I. Network construction and fundamental results. Phys. Rev. E 64 (1), 016131.
Newman, M.E., 2001b. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Phys. Rev. E 64 (1), 016132.
Newman, M.E., 2004. Coauthorship networks and patterns of scientific collaboration. Proc. Natl. Acad. Sci. 101 (suppl 1), 5200–5205.
Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The Pagerank Citation Ranking: Bringing Order to the Web Technical Report. Stanford University Database Group.
Strotmann, A., Zhao, D., Bubela, T., 2009. Author name disambiguation for collaboration network analysis and visualization. Proc. Am. Soc. Inform. Sci. Technol. 46 (1), 1–20.
Sun, L., Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling. Transport. Res. Part C: Emerg. Technol. 77, 49–66.
Wagner, C.S., Leydesdorff, L., 2005. Network structure, self-organization, and the growth of international collaboration in science. Res. Policy 34 (10), 1608–1618.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press.
Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 431316 (5827), 1036–1039.
Yan, E., Ding, Y., 2009. Applying centrality measures to impact analysis: a coauthorship network analysis. J. Am. Soc. Inform. Sci. Technol. 60 (10), 2107–2118.